DOCUMENT RESUME

ED 274 672 TM 860 524

AUTHOR Bangert-Drowns, Robert L.; And Others
TITLE Effects of Frequent Classroom Testing.
PUB DATE Apr 86
NOTE 45p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; *Classroom Techniques; Correlation; Criterion Referenced Tests; *Effect Size; Higher Education; High Schools; *Mastery Tests; *Meta Analysis; Student Attitudes; *Testing
IDENTIFIERS *Adjunct Questions

ABSTRACT

The educational effects of frequent classroom testing have been studied and discussed since the early part of this century. Many possible reasons for its presumed benefits have been offered, such as increased practice, review, improved course organization, and frequent opportunities for feedback. However, reviewers of evaluations of frequent testing have been unsure about whether the benefits are real. The present review makes two contributions to this discussion. First, it distinguishes the research on frequent classroom testing from two related areas, research on adjunct questions and research on mastery testing. Second, it provides a meta-analytic review of research on frequent testing. The meta-analysis shows that tested students performed about a half standard deviation better than untested students on summative measures. However, improvement was much smaller, about one-tenth of a standared deviation, when frequently tested students were compared to other students who also received tests, only less frequently. Frequently tested students rated their classes more favorably than did students who were less frequently tested. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED274672

Effects of Frequent Classroom Testing

Robert L. Bangert-Drowns, James A. Kulik,

& Chen-Lin C. Kulik

The University of Michigan

Based on a paper presentation at the annual meeting

of the American Educational Research Association,

San Francisco, April, 1986

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Bangert-Drowns

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

TM 860 524

Abstract

The educational effects of frequent classroom testing have been studied and discussed since the early part of this century. Many possible reasons for its presumed benefits have been offered, such as, increased practice, review, improved course organization, and frequent opportunities for feedback. However, reviewers of evaluations of frequent testing have been unsure about whether the benefits are real. The present review makes two contributions to this discussion. First, it distinguishes the research on frequent classroom testing from two related areas, research on adjunct questions and research on mastery testing. Second, it provides a meta-analytic review of research on frequent testing. The meta-analysis shows that tested students performed about a half standard deviation better than untested students on summative measures. However, improvement was much smaller, about one-tenth of a standard deviation, when frequently tested students were compared to other students who also received tests, only less frequently.

Effects of Frequent Classroom Testing

Written classroom tests were not always common features of school instruction. Prior to the middle of the nineteenth century, written materials were less available, and teachers used oral recitation to make students memorize course material and to assess students' educational success (Lindvall, 1961). The written essay test slowly began to replace recitation but it was not until the new century that techniques were developed to allow the proliferation of classroom testing (Fenton, 1929). To assess the aptitudes of large groups of men during the First World War, tests composed of "new type" or objective items were constructed. Objective tests were not as cumbersome to score as essays or individual performance tests and scores on objective tests could be explicitly calculated by cumulating performance on the items. As methods for constructing objective items filtered through the educational system, it became easy for teachers to test their students more frequently.

Almost immediately, educators and researchers began to wonder if such frequent testing could be more than just a means of assessment but have real educational value in itself. There are many good reasons for expecting frequent testing to increase student achievement. Frequent testing may encourage students to study more frequently, reduce their anxiety about tests, provide opportunities for reward and feedback about performance, stimulate mental review processes, indicate the teacher's expectations for course outcomes, and divide a course into smaller, more manageable units. Unfortunately, it is also easy to think that frequent testing might hinder student performance. Frequent testing may take time away from valuable regular instruction, encourage more attention to

tests than to course content, inhibit integration of larger units of instructional material, become tedious for students, or reduce student motivation to do well on the final exam.

Despite the decades of speculation about the effects of classroom testing, research in this area has been inconclusive. A few research areas, such as the research on adjunct questions or testing for mastery, are suggestive of the importance of testing yet are distinguishable from studies of testing in conventional settings. Those reviews that specifically looked at the effects of testing in conventional classes (Ross & Stanley, 1954; Proger & Mann, 1973) were done before the present standards of review methodology were developed and their findings are contradictory.

This article will repair the deficiency in two ways. First, we will briefly report findings from two related areas, adjunct questions and testing for mastery, to illuminate the research on classroom testing. Second, we will report the results of a meta-analysis of classroom testing.

Adjunct questions

In typical studies of adjunct questions, textual material was punctuated at intervals of 2 or 3 pages by "test-like questions" (Rothkopf, 1966) covering content that immediately preceded or would immediately follow the questions. Researchers used a criterion test to compare the performance of students who read the text with or without the test-like adjunct questions. By varying qualities of the texts, the adjunct questions, and the criterion test, researchers could test hypotheses about the types of cognitive processes that were being affected.

McGaw and Grotelueshen (1972) reported bidirectional effects for adjunct questions. To demonstrate a backward processing effect, criterion items were created from the same paragraphs from which the adjunct items were taken. Experimental subjects outperformed control subjects on these "matched" criterion items, but only on matched items from paragraphs that immediately preceded adjunct questions. The authors thought that adjunct questions activated a process of mental review, thereby increasing the retention of preceding information. They explained that this backward processing effect was most powerful for content located nearest adjunct questions.

McGaw and Grotelueshen (1972) also detected evidence in their data for a forward processing effect. Some items in their criterion measure came from the same page as the adjunct questions, but different paragraphs. On these "unmatched" items, the experimental subjects outperformed the control subjects only on criterion items that came from paragraphs immediately after adjunct questions. The authors proposed that adjunct questions heightened readers' attention and that this heightened attention increased the retention of material read immediately after the adjunct questions.

Nungester and Duchastel (1982) extended the design of adjunct question research to examine student performance over a longer unit of text. All subjects read a 1,700 word passage without interruption. Afterward, some subjects received an immediate posttest, a second group was asked to review the passage, and the other subjects completed a questionnaire as a filler task. Two weeks later all subjects received a retention test. Since the tested subjects outperformed subjects in the other two conditions, the authors advised the use of tests as an

instructional replacement for review. This result replicates findings of earlier researchers (e.g., Jones, 1923; Spitzer, 1939).

Rothkopf (1966) called adjunct questions "test-like events." Rickards (1979) confirmed and elaborated the bidirectional effects of adjunct questions in a broad review of relevant research. Nungester and Duchastel (1982) showed that tests given after texts are more instructionally beneficial than simple review or no tests. Can we generalize from these findings and expect that frequent classroom testing will stimulate the review and attention of the learner and thereby increase his or her achievement?

In all research, one must make a trade-off between internal and external validity. The design of adjunct question research exhibits a degree of control which allows for analytic study, but is unlike real school settings (Rickards, 1979; Duchastel, 1979). The positive effects of small sets of questions on the learning of short passages is established by these experiments, but the findings are only suggestive of what may result from real classroom testing.

Testing for mastery

Bloom (1968) and Keller (1968) advocated a particular use of frequent classroom testing. They assumed that, given sufficient time and quality of instruction, all students would be able to master specified educational content. In order to make this model workable, frequent tests are needed to serve as gateways for course progress. Students demonstrating mastery may progress to new instructional units. Students that did not achieve mastery on formative tests can be given detailed diagnosis and some kind of remediation on the basis of their performance.

There is a large body of research to show that testing for mastery is effective in increasing student achievement (Block & Burns, 1976; J. Kulik, Kulik, & Cohen, 1979; C. Kulik, Kulik, & Bangert-Drowns, 1986). But mastery testing obviously includes more than just frequent testing. At minimum, a mastery requirement (feedback and some kind of remediation) is added to frequent testing procedures. A mastery testing strategy may also include such features as self-pacing, use of tutors and an abandonment of conventional grades.

Some attempts have been made to partial out the effects of testing frequency from other features of mastery testing. Johnson and Ruskin (1977) located four studies that compared different frequencies of mastery testing. Three studies indicated improved performance with more testing; the fourth study showed no difference between conditions.

More recently, C. Kulik, Kulik, and Bangert-Drowns (1986) conducted a meta-analysis of 34 studies comparing mastery testing strategies to conventional testing strategies. Twenty-six studies compared two groups with the same number of tests, but only one of the groups had a mastery requirement. In the other eight studies, the two groups differed in both mastery requirement and frequency of testing. In all eight studies, the groups requiring mastery also had more frequent testing. The effect size for studies controlling test frequency was 0.45 standard deviations. Studies with mastery groups that also had more tests than the conventionally tested groups showed an average effect size of 0.74 standard deviations.

Clearly the evidence shows that when mastery is required on test performance, it is most effective to test frequently. However, in such courses, frequent testing is synonymous with smaller units to be

mastered and a greater likelihood for diagnosis and remediation of learning problems. This is not the case for conventional testing. The testing for mastery literature suggests the superiority of frequent testing, but this suggestion may not fit the conditions of the typical classroom test.

Frequent classroom testing

Almost as soon as objective testing was incorporated into the classroom, researchers began to look for its educational effects. Spitzer (1939), for example, referred to early studies conducted in 1914 and 1922 demonstrating that immediate written recall improved retention of newly learned material. Jones (1923) described a program of experiments determining the use of brief examinations to increase retention of information given in lectures. Turney (1931) and Kulp (1933) divided their students into two groups according to their performance on an exam. In both studies, the half that performed most poorly was given weekly tests in addition to regular exams; the other half received only the regular exams. The students that had previously performed poorly did as well on the final as their other classmates.

These early researchers believed that tests inhibited the normal decay of memory by rehearsal and "overlearning." Their inquiries set the stage, however, for a long history of speculation about the educational potency of testing. Jones (1923) suggested that testing operates as active practice, strengthening the associations between "total stimulus series" and "corresponding motor responses." McKeachie (1963) suspected that the ability of tests to provide knowledge of results is a key element of their efficacy. Rickards (1979) showed that even without providing feedback testing can affect kinds of cognitive

processing that facilitate retrieval of learned information. Mawhinney, Bostrow, Laws, Blumenfeld, & Hopkins (1971) demonstrated that students with a regime of daily testing studied more consistently than students with other patterns of testing. Ruskin and Johnson (1977) stressed that "learning is maximized when the learner is actively responding to the material to be learned" (p. 12). Classroom testing is a simple way to increase the frequency of student responses in instruction.

Reviewers have not been convinced that frequent classroom testing is an important instructional element. Ross and Stanley (1954) reported results of 16 studies of frequent classroom testing. Eight of these indicated that frequent testing had positive effects on student achievement measured immediately after the course. However, two studies had negative findings and six studies reported either mixed effects or no effects. In 1973, Proger and Mann conducted an extensive review of research on the effects of testing. They informally summarized evidence about frequent testing from 27 studies of varying quality. They concluded that the reports were too contradictory to demonstrate any improved achievement from increased testing frequency.

To date, no one has used quantitative review methods to summarize research on classroom testing. We chose to apply study effect meta-analysis (Bangert-Drowns, 1986). Study effects meta-analysis is a quantitative literature review that treats each study as a data point and uses conventional statistical tests to describe relations between study features and study findings. Our questions were straightforward. Does research show that frequency of testing affects student achievement? How large are these effects? Do the effects of frequent testing vary systematically according to features of the studies?

Method

Study effect meta-analysis is a modification of the approach described by Glass, McGaw, and Smith (1981). Like Glassian meta-analysis, study effect meta-analysis requires a reviewer (a) to locate studies of an issue through objective and replicable searches; (b) to code the studies for salient features; (c) to describe study outcomes on a common scale; and (d) to use statistical methods to find relations between study features and study outcomes.

Data Sources

To find studies on the effects of frequent testing, we carried out computer searches of two library data bases: (a) ERIC, a data base on educational materials from the Educational Resources Information Center, consisting of the two files Research in Education and Current Index to Journals in Education; and (b) Comprehensive Dissertation Abstracts. The empirical studies retrieved in these computer searches were the primary source of data for our analyses. A second source of data was a supplementary set of studies located by branching from bibliographies in the review articles retrieved by computer.

Search procedures yielded 36 studies that met five criteria for inclusion in our data set. First, the studies had to take place in real classrooms. Second, they had to compare two groups that received identical instruction except that one was tested more frequently than the other. Third, tests given to both groups were conventional classroom tests. Studies of mastery testing, for example, or studies of adjunct questions added to texts were not included in our final collection. Lastly, studies that had serious methodological flaws, such

as significant pretreatment differences between groups, and studies that were artificial, such as those using paid volunteers, were not included.

Study Features

Sixteen variables were used to describe treatments, methodologies, settings, and publication histories of the studies. These 16 variables were chosen on the basis of (a) an examination of variables used to describe study features in previous reviews, and (b) a preliminary examination of dimensions of variation in the studies located for this analysis. Two coders independently coded each of the studies on each of the variables. The coders then jointly reviewed their coding forms and discussed any disagreements. They resolved these disagreements by jointly reexamining the studies whose coding was in dispute.

Four of the sixteen variables described aspects of the testing procedures:

1. Number of tests given to the experimental group. This is the simple sum of the number of tests given to the frequently tested group during the study.
2. Number of tests given to the control group. Simply the number of tests given to the least frequently tested group in the study.
3. Feedback on tests. This feature distinguishes between studies in which students are given feedback about their test performance and studies where such information is not given.
4. Duration of treatment. Study duration was recorded in number of weeks.

Seven variables were used to describe the experimental designs of the studies:

1. Subject assignment. Students were assigned to experimental and control groups either randomly or by nonrandom procedures.
2. Control for teacher effects. Studies classified as having this control were those in which the same instructor or instructors taught both experimental and control groups. Studies classified as lacking this control were those in which different instructors taught experimental and control groups.
3. Control for historical effects. In studies classified as having this control, experimental and control groups were taught concurrently, e.g., in the same semester. In studies without this control, experimental and control groups were taught consecutively, e.g., in two different semesters.
4. Control for author-bias in criterion test. Studies classified as having a control for bias in test authorship were those that used standardized, commercial tests as criterion measures. Studies lacking this control used as criterion measures either locally developed tests or a combination of locally developed and commercial tests.
5. Control for bias in scoring the criterion test. Studies classified as having this control were those that used objective, machine-scorable examinations; studies without this control measured student achievement on essay tests or other nonobjective tests.
6. Amount of statistical control in outcome measurement. Outcomes were measured on a posttest alone in some studies, whereas in other studies outcomes were measured as pretest-posttest differences or as covariance-adjusted posttest scores.

7. Control for evaluator involvement. Studies classified as having this control were those in which the evaluator did not write teaching materials or teach the experimental or control groups. In studies lacking this control, the evaluator or study author wrote teaching materials or taught the experimental or control groups.

Three variables were used to describe features of the settings in which the evaluations were conducted:

1. Class level. Courses were at the precollege level or college level.
2. Course content. The subject matter taught in the courses was either (a) mathematics, (b) science, or (c) social sciences.
3. Student ability. Students were either of high, mixed, or low ability.

Two variables were used to describe the publication histories of the studies:

1. Year of the report. The publication or release year of each study was recorded.
2. Source of the study. The three document types were (a) technical reports, including clearinghouse documents, papers presented at conventions, etc.; (b) dissertations; and (c) professional publications, including articles, scholarly books, etc.

Outcome measures

The instructional outcome measured in 32 of the 36 studies was student learning, as indicated on achievement examinations given at the end of instruction. However, two other outcome variables were also considered.

Instead of measuring student learning through a summative achievement test, seven studies compared the performance of experimental and control groups on identical items received throughout the course. The items were distributed in short, frequent quizzes in the experimental treatment; in the control, the same items were massed in longer, less frequent tests. Another four studies measured student attitudes toward instruction.

For statistical analysis, outcomes had to be expressed on a common scale of measurement. The transformation used for this purpose was the one recommended by Glass et al. (1981). We coded each outcome as an effect size, defined as the difference between the mean scores of two groups divided by the standard deviation of the control group. For most studies, effect sizes could be calculated directly from reported means and standard deviations. For some studies, however, effect sizes had to be retrieved from *t* and *F* ratios. Formulas used in estimating effect sizes from such statistics were those given by Glass et al. (1981).

The application of the formulas given by Glass and his colleagues was straightforward in most cases. In some studies, however, more than one value was available for use in the numerator of the formula for calculating effect size, and more than one value was available for the denominator. For example, some investigators reported raw-score differences between groups as well as covariance-adjusted differences, and some reported differences on a post-measure as well as differences in pre-post gains. In such cases, we used as the numerator in the effect-size formula the difference that was least likely to be affected by individual differences among subjects and by other irrelevant factors. This meant using covariance-adjusted differences rather than

raw-score differences, and differences in gains rather than differences on posttests alone. In addition, some reports contained several measures of variation that might be considered for use as the denominator in the formula for calculating effect size. We used the measure that provided the best estimate of the unrestricted population variation in the criterion variable.

Unit of statistical analysis

Some studies reported more than one finding for a given outcome area. Such findings sometimes resulted from the use of more than one experimental or control group in a single study, and they sometimes resulted from the use of several subscales and subgroups to measure a single outcome. Using several effect sizes to represent results from one outcome area of one study seemed to be inappropriate to us because the effect sizes were usually nonindependent. They often came from a single group of subjects or from overlapping subject groups, and they almost always represented the effects of a single program implemented in a single setting. To represent a single outcome by several effect sizes would violate the assumption of independence necessary for many statistical tests and would also give undue weight to studies with multiple groups and multiple scales.

The procedure that we adopted, therefore, was to calculate only one effect size for each outcome area of each study. A single rule helped us to decide which effect size best represented the study's findings. The rule was to use the effect size from what would ordinarily be considered the most methodologically sound comparison when comparisons differed in methodological adequacy. (a) When results from both a true experimental comparison and a quasi-experiment were available from the

same study, results of the true experiment were recorded. (b) When results from a long and short implementation were available, results from the longer implementation were used. (c) When transfer effects were measured in addition to effects in the area of instruction, the direct effects were coded for the analysis. (d) In all other cases, our procedure was to use total scores and total group results rather than subscore and subgroup results in calculating effect sizes.

Results

Because 32 of the 36 studies in the pool investigated the effects of frequent classroom testing on summative examination performance, we were able to carry out a complete statistical analysis of results in this area. The analysis covered both average effects and the relationship between study effect and study features. We carried out less complete statistical analyses of other outcome areas because of the limited number of studies in these areas.

Examination performance

Thirty-two studies on frequent testing reported results from examinations given at the end of instruction (Table 1). Twenty-six of these studies showed results favoring the use of frequent testing. Six studies showed negative effects for frequent testing. Eleven of 26 studies with positive findings reported that the difference in amount learned by experimental and control groups was great enough to be considered statistically significant. One of the negative reports was statistically significant. Judging from the box-score results, one would conclude that frequent classroom testing is beneficial to student achievement.

Insert Table 1 about here

The index of effect size provides a more precise measure of the strength of the treatment effects. The average of the 32 effect sizes was 0.20. That is, the average effect of frequent testing was to raise achievement scores by 0.20 standard deviations. The standard error of the mean was 0.06. This effect is significant by conventional statistical standards, $t(31) = 3.33$, $p < .01$. An alternative representation of this size of effect can be taken from areas of the standard normal curve. Approximately 58% of the area of this curve falls below a z-score of 0.20. Therefore, students from frequently tested classes performed at the 58th percentile on their examinations, whereas students who were taught with less frequent testing performed at the 50th percentile on the same examinations. Put another way, 58% of the students who were frequently tested outperformed the average student who was not frequently tested.

On examining these studies, a complication became immediately apparent. Frequent testing was not consistently defined across studies. This meant, for example, that the administration of four tests was the frequent testing treatment in Wiggins (1968), but was the least frequent testing treatment in Noll (1939). It is hard to know how such contradictory definitions affect conclusions drawn from this literature.

To remedy the problem, we defined minimum and maximum rates for frequency of testing in experimental and control treatments respectively. Experimental groups had to receive tests at least every two weeks to be included. Control groups could not receive more than monthly testing to be included. Twenty-eight studies met these criteria

and constituted our selective sample. Twenty-four of these reported results from examinations given at the end of instruction. Further analyses were done on studies from the selective sample.

The average effect size of the 24 studies in the selective sample was 0.26 standard deviations. The standard error of the mean was 0.05. Again, this effect was significant, $t(23) = 5.2$, $p < .001$. According to data from the restricted sample, frequently tested students performed at the 60th percentile on their examinations; less frequently tested students at the 50th percentile.

Examination performance and study features

Our selective pool of studies on frequent testing showed a good deal of variance in their outcomes. Wiggins (1968) reported the largest effect, 0.79 standard deviations. At the other extreme, Nation, Knight, Lamberth, and Dyck (1974) reported a decrease of -0.22 standard deviations in examination scores after a treatment of frequent testing. If this variation in study outcome is systematic, it might be explainable in terms of variations in study features. We began further analyses to see if different kinds of studies produced different results (Table 2).

Insert Table 2 about here

Testing frequency in the control condition was the most important predictor of effect size. It is possible to be even more specific. Effect sizes were largest when the frequently tested group was compared to a control group that received no test. For studies where there was no control test, the average effect size was 0.46. When the control

students received one or more tests, the effect size dropped to an average of 0.13.

It is important to note that while the number of tests given to the control was a good predictor of effect size, the number of tests given to the frequently tested group was not related to effect size. The correlation between effect size and the number of tests given to the frequently tested group was -.06. This was an unexpected result. It indicates that as the frequency of testing increases in the frequently tested classes, there is no corresponding increase in student achievement.

We attempted to confirm this finding by further analyses. For example, if a simple count of number of tests is too crude a description of the treatment, a measure of test density might be more powerful. We regrouped studies by number of tests per week. The frequently tested group either received more than, less than, or exactly one test each week. When the studies were grouped into these categories, they still showed no relation to effect size ($r = .12$). Treatments receiving less than one test each week have an average effect size of 0.27, those with more than one test 0.33, and those with exactly one test 0.18.

As another confirmation of this finding, we collected all studies in the total sample in which the control group received no tests. In these ten studies, we calculated the correlation between effect size and the number of tests given to the experimental group. When holding the number of tests given to the control group constant at zero, frequency of testing in the experimental treatment correlated -.22 with effect size. A correlation of this size is not statistically significant ($p > .05$).

There was one piece of evidence that suggested frequency of testing does affect student achievement (Table 3). We identified eight studies that compared a high and intermediate frequency of testing with a low frequency. In seven of the eight studies, the highest frequency had a larger effect than the intermediate frequency.

Insert Table 3 about here

These relations are clearer when expressed in effect sizes. An average effect size was obtained for comparisons of the highest and lowest testing frequency in the eight studies. The effect size was 0.48. The average effect size for comparisons between the intermediate and lowest testing frequency was 0.22. The difference between these two, 0.26, was statistically significant, $t(7) = 3.25$, $p < 0.05$. In these eight studies, therefore, more frequent testing was superior to lesser frequencies in improving student achievement.

Attitudes toward instruction

Four studies measured students' attitudes toward instruction after they were exposed to differing testing conditions (Table 4). In all four studies, frequently tested students rated their classes more favorably than did students who were less frequently tested. The average of effect sizes drawn from these four studies was 0.59 standard deviations. This is a large effect.

Insert Table 4 about here

Performance on differentially distributed test items

Seven studies reported the performance of students on the test items given during instruction (Table 5). In these seven studies, experimental and control students were tested with the same items. In the experimental classes, the items were distributed in short, frequent quizzes; in the control, the identical items were given in longer, less frequent tests. In the typical study, the experimental group outperformed the control group by 0.49 standard deviations. Students perform better on short quizzes on short units of instruction than on longer quizzes over longer units of instruction.

Insert Table 5 about here

Discussion

Since the second decade of this century, teachers and researchers have speculated about and studied the effects of frequent classroom testing. Despite their efforts, the evidence on frequent testing has not returned a verdict. While early proponents (e.g., Jones, 1923) expected great benefits from classroom testing, researchers such as Noll (1939) made early complaints of negative effects. Narrative reviews of the literature (Ross & Stanley, 1954; Proger & Mann, 1973) were inconclusive because they could not reconcile the apparent contradictions in research findings.

Certainly when one looks at research in related fields, one is led to expect positive effects from frequent classroom testing. Findings from adjunct question research (Rickards, 1979) consistently showed that dividing texts into small units with "test-like events" improved student achievement. Research on mastery testing indicated that frequent

testing, at least when used as a diagnostic tool and followed with some kind of remediation, would improve classroom learning (Johnson & Ruskin, 1977; C. Kulik, Kulik, & Bangert-Drowns, 1986).

However, the typical settings and testing conditions in which these other methods have demonstrated their effectiveness are very different from those of classroom testing. In the classroom, information to be learned is presented with a good deal of redundancy. Typically, the student is given the same information through test, lecture, exercises, and media presentations. The student is generally encouraged to review this material before a test; the review may even be led by the teacher. Finally, the student is aware that his or her test performance is a "one-time" event that will permanently contribute to his or her academic record.

These characteristics of classroom tests make them very different from repeated testing, adjunct questions, and mastery testing as studied in the literature. There is no classroom instruction in the repeated testing studies nor opportunities for pretest review in adjunct question research. Classroom instruction generally presents information in larger units more redundantly and through more media than in the typical adjunct question research. As for mastery testing, there is less emphasis on grades and more on diagnosis and remediation than in typical classroom settings.

Despite the differences between research on classroom testing and on other uses of tests, this meta-analysis produced findings that are not inconsistent with the other literatures. We found, for example, that the presence of classroom testing does increase performance on summative measures of achievement. When tested groups were compared to

untested groups, the tested groups typically scored about a half standard deviation better on a summative criterion than did the untested students.

However, the meta-analysis did not produce findings to indicate that frequency of testing significantly influenced academic achievement. That is, when tested students were compared to other, less frequently tested students, they only showed minimal, nonsignificant superiority on a summative achievement measure, about one tenth of a standard deviation. The correlation between the number of tests given to the frequently tested students and effect size was actually negative, r = -.06, though not significantly so. This absence of relation between frequency of testing and effect size remained even when we considered only those studies in which the control group received no tests.

We did find that when two groups were given identical test items, superior performance was obtained from students that received the items more frequently in shorter tests on smaller units of instructional content. In fact, the more frequently tested group performed 0.49 standard deviations better than students in the comparison group. This, of course, is not evidence that frequent testing improves performance on summative measures, only on test items distributed differentially through the course.

There was one bit of contradictory evidence. We found eight studies that not only compared frequent and infrequent testing, but had a third intermediate testing condition. In seven of the eight, greater student achievement was systematically related to frequencies of testing; higher frequencies produced higher achievement. This evidence must be taken seriously. Although meta-analysis is helpful in

summarizing findings in a diverse literature, the kind of "internal" evidence in these eight studies is better controlled. Findings within each study come from the same course in the same educational setting and were collected by the same researcher. Such variables can only be controlled within studies, not between. We conclude, therefore, that frequent classroom testing was systematically related to summative achievement in these eight studies. It remains unclear, however, whether one can speak generally about the effects of frequent testing on the basis of evidence from these eight studies.

There is no characteristic that the eight studies have in common nor are they distinguishable from the other studies on the basis of some study feature. However, the eight studies are unlike the others in their average effect size. In the whole group of studies, the average effect of classroom testing on achievement was 0.20 standard deviations. In the subsample of eight studies, the average effect of frequent testing (comparing the conditions of most and least frequency) was 0.48 standard deviations, quite a bit larger than the overall average. Thus, these eight studies cannot be taken as representative of the research on frequent testing as a whole. Either they share a special feature that was not reported or the occurrence of this group of eight studies was a chance event.

Despite the failure of the research to show that frequency of testing affects academic achievement, there was evidence that frequent testing was related to an important affective outcome. Five studies measured students' attitude toward instruction following programs of varying test frequency. The frequent testing condition had the effect of making students' attitudes more positive by 0.46 standard deviations.

That is, students in these studies had a more favorable opinion of their instruction when they were tested frequently. Increasing the frequency of tests may be a way of creating a more positive atmosphere in the classroom.

REFERENCES

Badia, P., Harsh, J., & Stutts, C. (1978). An assessment of methods of instruction and measures of ability. Journal of Personalized Instruction, 3, 69-75.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. Psychological Bulletin, 99.

Block, J. H., & Burns, R. B. (1976). Mastery learning. In L. S. Shulman (Ed.), Review of Research in Education, (Vol. 4). Itasca, IL: F. E. Peacock.

Bloom, B. S. (1968 May). Mastery learning. Evaluation Comment, (Vol. 1, No. 2). Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation of Instructional Programs.

Corbett, J. S., Jr. (1978). Mastery of course material as a function of workbook assignments and test frequency. Masters Abstracts, 17, 33. (University Microfilms No. 1311964)

Curo, D. M. (1963). An investigation of the influence of daily pre-class testing on achievement in high school American history classes. Dissertation Abstracts International, 24, 5236. (University Microfilms No. 64-4574)

Deputy, E. C. (1929). Knowledge of success as motivating influence in college work. Journal of Educational Research, 20, 327-334.

Duchastel, P. C. (1979). Adjunct questions effects and experimental constraints. Occasional paper 1. Bryn Mawr, PA: American College. (ERIC Document Reproduction Service No. ED 216 312)

Dustin, D. (1971). Some effects of exam frequency. The Psychological Record, 21, 409-414.

Fenton, N. (1929). New type examinations and their daily use in the classroom. Education, 50, 150-158.

Fitch, M. L., Drucker, A. J., & Norton, J. A., Jr. (1951). Frequent testing as a motivating factor in large lecture classes. Journal of Educational Psychology, 42, 1-20.

Fulkerson, F. E., & Martin, G. (1981). Effects of exam frequency on student performance, evaluations of instructor, and test anxiety. Teaching of Psychology, 8, 90-93.

Gable, F. (1936). The effect of two contrasting forms of testing upon learning. Baltimore: Johns Hopkins University.

Gaynor, J., & Willham, J. (1976). Student performance and evaluation under variant teaching and testing methods in a large college course. Journal of Educational Psychology, 68, 312-317.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage Publications.

Johnson, K. R., & Ruskin, R. S. (1977). Behavioral instruction: An evaluative review. Washington, D.C.: American Psychological Association, Inc.

Jones, H. E. (1923). Experimental studies of college teaching: The effect of examination on permanence of learning. Archives of Psychology, 10, 1-70.

Keller, F. S. (1968). "Good-bye, teacher..." Journal of Applied Behavioral Analysis, 1, 79-89.

Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. Journal of Educational Psychology, 25, 427-436.

Kirkpatrick, J. E. (1934). The motivating effect of a specific type of testing program. (Doctoral dissertation, Iowa University, 1933). University of Iowa Studies in Education, 9, 41-68.

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1986). Effects of testing for mastery: A meta-analysis. Manuscript submitted for publication.

Kulik, J. A., Kulik, C.-L., & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. American Psychologist, 34, 307-318.

Kulp, D. H., II (1933). Weekly tests for graduate students? School and Society, 38, 157-159.

Laidlaw, W. J. (1963). The effects of frequent tests on achievement, retention, and transfer, and test behavior. Dissertation Abstracts International, 24, 5197. (University Microfilms No. 64-4322)

Lindenberg, T. S. (1984). The effect of test frequency on achievement in the first principles of accounting course. Dissertation Abstracts International, 45, 2736A. (University Microfilms No. DA8426700)

Lindvall, C. M. (1961). Testing and evaluation: An introduction. New York: Harcourt, Brace, & World, Inc.

Mach, G. R., Jr. (1963). A comparative study of student performance in an intermediate calculus class as a result of different evaluation programs. Dissertation Abstracts International, 24, 5248. (University Microfilms No. 64-5745)

Maloney, E. L., & Ruch, G. M. (1929). The use of objective tests in teaching as illustrated by grammar. School Review, 37, 62-66.

Marso, R. N. (1970). Classroom testing procedures, test anxiety, and achievement. Journal of Experimental Education, 38, 54-58.

Mawhinney, V. T., Bostrow, D. E., Laws, D. R., Blumenfeld, G. J., and Hopkins, B. L. (1971). A comparison of students studying-behavior produced by daily, weekly , and three-week testing schedules. Journal of Applied Behavior Analysis, 4, 257-264.

McGaw, G., & Grotelueschen, A. (1972). Direction of the effect of questions in prose material. Journal of Educational Psychology, 63, 580-588.

McKeachie, W. J. (1963). Research on teaching at the college and university level. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally.

Monk, J. J., & Stallings, W. M. (1971). Another look at the relationship between frequency of testing and learning. Science Education, 55, 183-188.

Mudgett, A. G. (1956). The effects of periodic testing on learning and retention in engineering drawing. Dissertation Abstracts International, 16, 2351-2352. (University Microfilms No. 56-3744)

Nation, J. R., Knight, J. M., Lamberth, J., & Dyck, D. G. (1974). Programmed student achievement: A test of the avoidance hypothesis. The Journal of Experimental Education, 42, 57-61.

Negin, G. A. (1981). The effects of test frequency in a first-year torts course. Journal of Legal Education, 31, 673-676.

Noll, V. H. (1939). The effect of written tests upon achievement in college classes: An experiment and a summary of evidence. Journal of Educational Research, 32, 345-358.

Nungester, R. J. and Duchastel, P. C. (1982). Test versus review: Effects on retention. Journal of Educational Psychology, 74, 18-22.

Nystrom, N. K. (1969). An experimental study to compare the relative effects of two methods of instruction on learning of intermediate algebra. Dissertation Abstracts International, 29, 3532A-3533A.

Palmer, E. L. (1974). Frequency of tests and general subject-area mastery. Psychological Reports, 35, 422.

Pikunas, J., & Mazzota, D. (1965). The effects of weekly testing in the teaching of science. Science Education, 49, 373-376.

Pratt, F. H. (1970). The effect of frequent testing in American History upon student achievement and attitudes. Dissertation Abstracts International, 31, 80A. (University Microfilms No. 70-11,892)

Proger, B. B., & Mann, L. (1973). An historical review of theoretical and experimental literature on the teaching values of informal (nonstandardized), teacher-made achievement tests: 1913-1968. Blue Bell, PA: Montgomery County Intermediate Unit 23. (ERIC Document Reproduction Service No. ED 084 292)

Rickards, J. (1979). Adjunct postquestions in text: A critical review of methods and processes. Review of Educational Research, 49, 181-196.

Rievman, S. P. (1973). Optimal frequency of testing as a function of ability level and reinforcement history. Dissertation Abstracts International, 34, 7054A. (University Microfilms No. 74-11,605)

Robinson, P. (1972). Contingent systems of instruction. Paper presented at the Rocky Mountain Psychological Association Convention. (ERIC Document Reproduction Service No. ED 069 704)

Ross, C. C., & Henry, L. K. (1939). The relation between frequency of testing and progress in learning psychology. Journal of Educational Psychology, 30, 604-611.

Ross, C. C., & Stanley, J. C. (1955). Measurements in Today's Schools. New York: Prentice-Hall, Inc.

Rothkopf, E. Z. (1966). Learning from written instruction material: An exploration of the control of inspection behavior by test-like events. American Educational Research Journal, 3, 241-249.

Selakovich, D. (1962). An experiment attempting to determine the effectiveness of frequent testing as an aid to learning in beginning college courses in American Government. The Journal of Educational Research, 55, 178-180.

Shapiro, S. L. (1973). An experimental study of the effects of frequency of testing procedures on students in a business organization and management course in a community college with an open admissions policy. Dissertation Abstracts International, 35, 4207A-4208. (University Microfilms No. 74-19,774)

Spitzer, H. F. (1939). Studies in retention. Journal of Educational Psychology, 30, 641-656.

Standlee, L. S., & Popham, W. J. (1960). Quizzes' contribution to learning. Journal of Educational Psychology, 51, 322-325.

Townsend, N. R., & Wheatley, G. H. (1975). Analysis of frequency of tests and varying feedback delays in college mathematics achievement. College Student Journal, 9, 32-36.

Turney, A. H. (1931). The effect of frequent short objective tests upon the achievement of college students in educational psychology. School and Society, 33, 760-762.

Vernon, P. E. (1954). Practice and coaching effects in intelligence tests. Educational Forum, 18, 269-280.

Ward, E. F. (1984). Statistics mastery: A novel approach. Teaching of Psychology, 11, 223-225.

Wiggins, J. A. (1968). Learning contingencies in the college classrooms: A pilot study. Final report. (ERIC Document Reproduction Service No. ED 024 314)

Wilkins, S. A. (1979). A study of the effects of various teaching-testing frequencies on cognitive gains. Dissertation Abstracts International, 40, 1977A. (University Microfilms No. 79 21993)

Williams, R. L., & Lawrence, J. (1974). The effects of frequency of quizzing in a lecture course. Paper presented at the Second National Conference on Research and Technology in Higher Education, Atlanta.

Table 1

Major Features and Achievement Effect Sizes in 32 Studies on Frequent Classroom Testing

BEST COPY AVAILABLE

Study	Place	Class Level	Course Content	Duration in Weeks	No. of Tests for X groups	No. of Tests for C Groups	ES
Curo, 1964**	Indiana	11th	Social Science	6	25	2	0.10
Deputy, 1929	State University of New York	College	Philosophy	6	12	0	0.61
Fitch, Drucker, & Norton, 1951	Purdue University	College	Government	15	15	4	0.26
Fulkerson & Martin, 1981**	Western Illinois University	College	Psych	12	8	4	0.07
Gable, 1936**	Maryland	9th	Science	7	21	5	-0.80
Keys, 1934	University of California	College	Psych	15	8	2	-0.01
Kirkpatrick, 1934	Iowa	High School	Science	18	20	2	0.31
Laidlaw, 1963	Fairleigh Dickinson University	College	Psych	16	16	4	-0.08
Lindenberg, 1984	Illinois Community Colleges	College	Accounting	17	12	2	0.01
Mach, 1963	California State Polytechnic College	College	Math	12	29	1	0.13*
Maloney & Ruch, 1929	California	9-11	Reading	10	5	0	0.59
Marso, 1970**	University of Nebraska	College	Psych	15	6	3	0.14
Monk & Stallings, 1971**	University of Illinois	College	Geography	15	10	6	0.07
Mudgett, 1956	University of Minnesota	College	Engineering	12	36	2	0.26
Nation, Knight, Lamberth, & Dyck, 1974	University of Oklahoma	College	Psych	8	8	1	-0.22
Negin, 1981**	Marquette University	College	Law	15	3	0	0.70
Noll, 1939**	Rhode Island State College	College	Psych	15	5	1	-0.27
Nystrom, 1969**	California Junior College	College	Math	15	50	5	0.31
Palmer, 1974	Davison College	College	Psych	10	6	0	0.55
Pikunas & Mazzota, 1965	Michigan	12th	Science	6	6	0	0.71
Pratt, 1970	Arizona	High School	Social science	9	11	0	0.19
Robinson, 1972	Brigham Young University	College	Psych	4	3	0	0.10
Rievman, 1974	Florida Atlantic University	College	Psych	16	10	2	0.34

Table 1 (continued)

Study	Place	Class Level	Course Content	Duration in Weeks	No. of Tests for X groups	No. of Tests for C Groups	ES
Ross & Henry, 1939	Iowa State University	College	Psych	12	10	1	0.06
Selakovich, 1962	West Texas State College	College	Government	15	15	3	0.08
Shapiro, 1973	New York Community College	College	Business	15	10	3	0.26
Standlee & Popham, 1960	Indiana University	College	Psych	15	14	1	0.26
Townsend & Wheatley, 1975	California State Polytechnic College	College	Math	12	49	1	0.45*
Ward, 1984	Western Illinois University	College	Statistics	15	13	2	-0.13*
Wiggins, 1968	University of North Carolina	College	Sociology	5	4	0	0.79
Wilkins, 1979	Louisiana Community College	College	Psych	12	11	0	0.30
Williams & Lawrence, 1974	Western Michigan University	College	Physiology	4	3	0	0.34

*Estimated on the basis of direction and statistical significance of reported values

**Study not included in the selective sample

Table 2

Means and Standard Errors of Achievement Effect Sizes for 24 Studies on Frequent Classroom Testing Classified by Study Feature

Study Feature	N	M	SE
Number of tests for X Group			
5 or less	4	0.46	0.15
6-10	7	0.25	0.12
11-15	8	0.20	0.08
16 or more	5	0.21	0.09
Number of tests for C groups*			
None	9	.46	0.08
One to three	13	.14	0.06
Four to six	2	.09	0.17
Feedback on tests			
No	4	.20	0.16
Yes	15	.29	0.07
Duration of treatment			
8 weeks or less	6	0.39	0.15
9 to 12 weeks	8	0.32	0.07
13 weeks or more	10	0.13	0.06
Subject assignment			
Random	3	0.33	0.13
Nonrandom	21	0.25	0.06
Control for instructor effect			
Same	18	0.21	0.05
Different	5	0.38	0.17

Table 2(continued)

Study Feature	N	M	SE
Control for historical effect			
Same semester	22	0.23	0.18
Different semester	2	0.52	0.05
Control for author bias of test			
Commercial	5	0.33	0.13
Local	19	0.24	0.06
Scoring bias of test			
Objective	21	0.25	0.06
Non objective	3	0.30	0.02
Statistical control in outcome measures			
Post scores only	11	0.31	0.10
Control for covariate	13	0.21	0.05
Control for evaluator involvement			
Involved	20	0.25	0.05
Not involved	4	0.29	0.20
Class level			
Pre-college	4	0.45	0.12
College	20	0.22	0.06
Course content			
Math	3	0.15	0.17
Science	3	0.45	0.13
Social science	13	0.20	0.07
Others	5	0.35	0.11

Table 2(continued)

Study Feature	N	M	SE
Year of report			
Up to 1955	6	0.30	0.11
1956-1965	6	0.23	0.09
1966-1975	9	0.31	0.09
1976-1985	3	0.06	0.12
Source of study			
Unpublished	3	.41	0.20
Dissertation	8	.18	0.05
Published	13	.27	0.08

*Statistically significant, $p < .001$

Table 3

Effect Sizes in 8 Studies Using High, Intermediate, and Low Frequency of Testing

Study	No. of Tests for High Frequency Condition	No. of Tests for Intermediate Frequency Condition	No. of Tests for Low Frequency Condition	ES for High vs. Low Frequency of Testing	ES for Intermediate vs. Low Frequency of Testing
Deputy, 1929	12	6	0	0.96	0.26
Mudgett, 1956	36	12	2	0.26	0.06
Negin, 1981	3	1	0	0.70	0.36
Palmer, 1974	6	3	0	0.55	0.18
Reivman, 1973	10	4	2	0.34	0.26
Shapiro, 1973	10	5	3	0.26	0.19
Townsend, & Wheatley 1975	48	3	1	0.45*	0.10*
Wilkins, 1979	11	3	0	0.30	0.38

*Estimated on the basis of direction and statistical significance of reported values.

Table 4

Attitude Effect Sizes in 4 Studies

on Frequent Classroom Testing

Study	ES
Fulkerson & Martin (1981)	0.80
Gaynor & Millham (1976)	0.49
Nystrom (1968)	0.31
Shapiro (1972)	0.74

Table 5

Features and Achievement Effect Sizes

in 7 Studies Comparing Different Distributions of Quiz Items

Study	Place	No. of Tests for X Group	No. of Tests for C Group	ES
Badia, 1978	Bowling Green State University	9	3	0.25
Dustin, 1971	SUNY, Plattsburg	4	1	0.62
Fulkerson & Martin, 1981	Western Illinois University	8	4	0.60
Gaynor & Milham, 1976	University of Houston	12	2	0.21
Keys, 1934	University of California	8	2	0.53
Monk & Stallings, 1971	University of Illinois	8	4	0.22
Rievman, 1973	Florida Atlantic University	10	2	1.02